[IR2018-19-HW1] 1183732 Davide Martini

Indice:

1. Descrizione del lavoro svolto

Lo sviluppo dell'homework è stato effettuato in ambiente *Linux, Ubuntu 18.04*, tramite l'utilizzo di:

- *terrier-core-4.4* per l'indicizzazione dei documenti e la generazione delle varie run [2].
- *trec_eval 9.0.4* per ottenere le misure di valutazione dei vari sistemi [3].
- *os* libreria *Python* per le chiamate di sistema [4].
- *matplotlib* libreria *Python* per la creazione dei grafici riassuntivi [5].
- *statsmodel* libreria *Python* per il calcolo del test ANOVA 1-way e il Tukey HSD test [6].
- *numpy, scipy* librerie *Python* accessorie [7].
- *PyCharm 2018.3* programma utilizzato per lo sviluppo ed il test dello script [8].

È stato inoltre prodotto uno script in *Python 3.7* per automatizzare l'esecuzione dell'indicizzazione, la produzione delle varie run, il calcolo delle misure, produrre il test statistico ANOVA 1-way, il Tukey HSD test e per ottenere i plot finali. Lo script è contenuto nel repository [1] con il nome di '*ir_script.py'*. Di seguito verrà riportata la struttura del codice con una breve spiegazione del suo contenuto.

Per automatizzare l'esecuzione di tutto il workflow necessario per ottenere i risultati, il codice esegue i seguenti passi:

1. Viene impostata una cartella principale dove è contenuta la collezione di documenti, gli eseguibili dei software *terrier* e *trec_eval* e dove verranno salvati i risultati ottenuti.
2. Viene inizializzato il software *terrier* e vengono modificate le proprietà per l'indicizzazione e per l'esecuzione delle run. Questo passaggio viene ripetuto per tutti i modelli da analizzare. Sono stati salvati i vari file degli indici (non caricati a causa delle dimensioni) e i file delle run nella cartella '*run/*'.
3. Viene eseguita la valutazione tramite il software *trec_eval* sui file risultanti delle run. I file contenenti le misure per la valutazione sono stati successivamente salvati nella cartella '*run/eval/*'.
4. È stata creata una struttura per contenere le varie misure di valutazione ottenute dal software *trec_eval*, in modo da avere un facile reperimento di queste per i test statistici e per i successivi plot.
5. Sono stati creati dei file, per ogni misura, in cui ogni colonna rappresenta una run.
6. È stata successivamente effettuata l'ANOVA 1-way ed il Tukey HSD test.
7. Sono stati infine prodotti i plot per ogni topic e run delle misure *P(10)* e *Rprec*. Un grafico contenente il valore della *MAP* per ogni modello analizzato.

Le impostazioni, utilizzate per l'analisi dei file TREC, sia per i documenti da indicizzare che per le query usate per lo sviluppo delle run, sono riportate in *Tabella 1*. È stata inoltre aggiunta la proprietà *ignore.low.idf.terms=true*, così da ignorare i termini con valore basso di IDF, in modo da non condizionare il reperimento di documenti con termini molto frequenti anche nei sistemi dove non viene considerata la stop list.

TrecDocTags.doctag=DOC	TrecDocTags.idtag=DOCNO	TrecDocTags.skip=DOCHDR	TrecDocTags.casesensitive=false
TrecQueryTags.doctag=TOP	TrecQueryTags.idtag=NUM	TrecQueryTags.process=TITLE, DESC	TrecQueryTags.skip=NARR

***Tabella 1**. Impostazioni aggiunte nel file 'terrier.properties'.*

Per quanto riguarda le query, si è deciso di considerare sia il titolo che la descrizione del topic, in quanto si è pensato che questa scelta desse un maggior contributo nel reperimento dei documenti rilevanti. Non si è usato però anche la narrazione per alleggerire il carico di lavoro durante l'esecuzione delle query.

2. Risultati ottenuti

In *Tabella 2* vengono riportati il numero di documenti indicizzati e il numero di parole presenti nel vocabolario, in modo da notare come vari l'utilizzo o meno della stop list e del Porter stemmer.

I vari sistemi con le diverse impostazioni, sono riportati con una sigla:

- *BM25*: modello BM25 con stop list e Porter stemmer.
- *BM25_stem*: modello BM25 senza stop list, con Porter stemmer.
- *TF_IDF*: modello TF_IDF con stop list e Porter stemmer.
- *TF_IDF_not*: modello TF_IDF senza stop list e senza Porter stemmer.

	TF_IDF	***BM25***	***BM25_stem***	***TF_IDF_not***
documenti indicizzati	528155	528155	528155	528155
dimensione del dizionario	738439	738439	738643	840517

***Tabella 2**. Statistiche dei vari indici*

È stato sviluppato il test statistico ANOVA 1-way, Tukey HSD pairwise test e Tukey HSD multiple comparison.

L'obiettivo del test statistico ANOVA 1-way è capire se i vari sistemi di reperimento analizzati avessero, oppure no, la stessa media. Nello sviluppo del test, il valore di soglia $\alpha = 0.05$ permette di rifiutare o meno l'ipotesi che i 4 sistemi abbiamo la stessa media. I valori ottenuti sono presenti nei file '*run/plot/anovaMEASURE.txt*', dove *MEASURE* indica la misura presa in considerazione, i quali vengono riportati in *Tabella 3*.

AP		***P(10)***		***Rprec***	
F_{stat}	~ 0.2698	F_{stat}	~ 0.3578	F_{stat}	~ 0.3508
$p = P[F \geq F_{stat} \mid H_0]$	~ 0.8471	$p = P[F \geq F_{stat} \mid H_0]$	~ 0.7836	$p = P[F \geq F_{stat} \mid H_0]$	~ 0.7886

***Tabella 3**. ANOVA 1-way.*

Si nota quindi che l'ipotesi non viene rifiutata e quindi i 4 sistemi di reperimento dell'informazione hanno la stessa media per le tre misure analizzate.

Tukey HSD, α = 0.05					
group 1	***group 2***	***meandiff***	***lower***	***upper***	***reject***
BM25	*BM25_stem*	-0.0018	-0.0877	0.0842	False
BM25	*TF_IDF*	-0.0005	-0.0865	0.0855	False
BM25	*TF_IDF_not*	-0.0251	-0.1111	0.0609	False
BM25_stem	*TF_IDF*	0.0012	-0.0848	0.0872	False
BM25_stem	*TF_IDF_not*	-0.0233	-0.1093	0.0626	False
TF_IDF	*TF_IDF_not*	-0.0246	-0.1106	0.0614	False

***Tabella 4**. Tukey HSD pairwise test - AP.*

Per quanto riguarda il Tukey HSD pairwise test, i risultati ottenuti sono riportati nella *Tabella 4* per la misura *AP*. Il test per le altre misure è presente nella cartella *'run/plot/TukeyHSDMEASURE.txt'*, dove *MEASURE* è la misura considerata.

Dal Tukey HSD test si nota che ogni coppia di sistemi è simile e c'è sempre intersezione negli intervalli di confidenza. Per avere una visione d'insieme è stato effettuato un confronto multiplo tra i vari modelli riportato in *Figura 1*, *Figura 2* e *Figura 3* e rispettivamente nei file '*run/plot/TukeyHSDtestMEASURE.svg*', con al posto di *MEASURE*, la misura di valutazione desiderata.
I sistemi adottati quindi sono molto simili, anche se l'indicizzazione del metodo *TF_IDF_not* è risultata più lunga in quanto sono state tenute in considerazione tutte le parole e queste non sono state sottoposte al processo di stemming.
Per quanto riguarda l'analisi della misura *Rprec* il confronto dei grafici, presenti nella cartella '*run/plot/*' con nome '*RprecSYSTEM.svg*', dove *SYSTEM* indica il sistema considerato, indica una lieve differenza tra i metodi *BM25*, *BM25_stem*, *TF_IDF*, ma evidenzia un forte calo in *TF_IDF_not*. Questa differenza non è presente in tutti i topic ma è abbastanza accentuata nei casi in cui questa si verifica.
I valori ottenuti per la misura *P(10)*, presenti nei grafici nella cartella '*run/plot*' con nome '*P_10SYSTEM.svg*' dove *SYSTEM* indica il sistema considerato, mostrano un andamento abbastanza simile per i metodi che adottano *BM25*, *BM25_stem*, *TF_IDF*, ma *TF_IDF* ha un andamento simile per alcuni topic e un valore minore della *Precision* per gli altri.
In *Figura 1*, *Figura 2*, *Figura 3*, sono riportati i test di Tukey per le misure *AP*, *P(10)* e *Rprec* rispettivamente. I test sono stati eseguiti per le tre misure, in quanto sia la *P(10)* che la *Rprec* presentavano valori simili per i sistemi *TF_IDF* e *BM25*. L'obiettivo del test era quello di capire se questi sistemi fossero oppure no simili. Dai grafici si nota che la null hypothesis non viene mai rifiutata, quindi non c'è una significativa differenza tra i sistemi di reperimento dell'informazione analizzati.

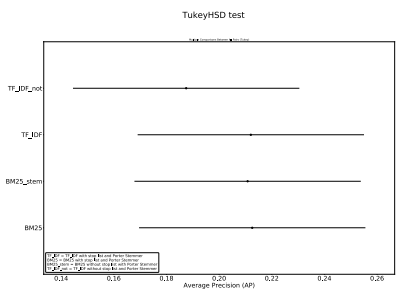

Figura 1. *Tukey HSD test – AP.*

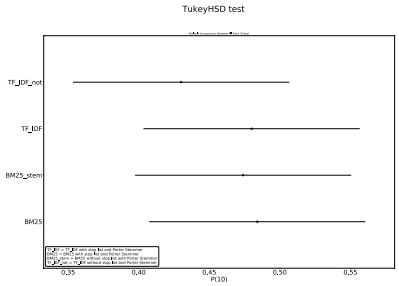

Figura 2. *Tukey HSD test – P(10).*

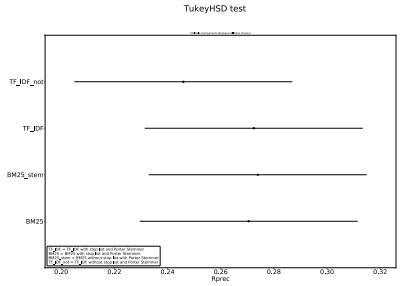

Figura 3. *Tukey HSD test – Rprec.*

Durante la fase di indicizzazione è stato conteggiato il tempo necessario per la sua esecuzione per ogni sistema, i quali vengono riportati in *Tabella 5*. I numeri rappresentano i secondi utilizzati per effettuare l'indicizzazione dei documenti.

Si può notare che ovviamente nei due sistemi nei quali non vengono eliminate le stop word il tempo di esecuzione è maggiore. Questo è un elemento che potrebbe incidere nella scelta di un metodo rispetto ad un altro.

TF_IDF	***BM25***	***BM25_stem***	***TF_IDF_not***
261.7906	261.7906	296.0097	296.6661

Tabella 5*. Tempo di indicizzazione della collezione (sec).*

3. Considerazioni finali

Per ottenere delle considerazioni finali su quale sia il sistema migliore tra quelli analizzati, quale sia il peggiore o quali differenze siano presenti tra i vari metodi, è stata analizzata sia l'efficienza che l'efficacia dei vari IRS. Questi due aspetti sono stati considerati su una collezione di documenti modesta e le considerazioni che emergono da questa analisi potrebbero essere molto diverse per certi aspetti ma per altri potrebbero essere un chiaro indicatore per sviluppare nuove analisi o scartare già a priori uno o più sistemi analizzati per collezioni di dimensioni maggiori.
Per quanto riguarda l'efficienza, è stato rilevato il tempo richiesto per l'indicizzazione della collezione nei vari casi. Questi valori sono stati riportati in *Tabella 5* e viene evidenziato un aumento del tempo richiesto di circa il 10% nei sistemi che non fanno uso della stop list. Questo valore è molto importante, in quanto per la collezione considerata, questa differenza è di circa 30 secondi, ma se venisse considerata una collezione con un numero di documenti molto maggiore questo aspetto potrebbe avere un forte impatto. L'utilizzo o meno di una stop list si rivela un fattore molto importante per l'efficienza dei sistemi, lo sviluppo di una fase di stemming, per quanto riguarda il tempo di esecuzione per la fase dell'indicizzazione, non si rivela così pesante. Possiamo quindi considerare, per quanto riguarda l'efficienza, solamente i metodi *TF_IDF* e *BM25*.
Per quanto riguarda l'efficacia dei vari sistemi, sono state svolte analisi statistiche sulle misure *AP*, *P(10)* e *Rprec*. Come riportato in *Tabella 3* e in *Figura (1-3)*, il test statistico ANOVA 1-way ed il Tukey HSD test mettono in evidenza che tra i 4 sistemi considerati non è presente nemmeno uno che abbia caratteristiche significativamente diverse rispetto agli altri. Il test ANOVA 1-way non rifiuta in nessun caso la null hypothesis e quindi vengono considerate le medie delle misure *AP*, *P(10)* e *Rprec* uguali nei sistemi considerati.
Il Tukey HSD test in *Figura (1-3)* mostra che ci sono delle differenze tra i vari metodi di reperimento dell'informazione utilizzati, ma è sempre presente una abbondante intersezione negli intervalli di confidenza per tutte le misure considerate.
Occorre però notare che i valori di *Rprec* e *P(10)* sono molto più bassi nel sistema *TF_IDF_not*, il quale ha inoltre un dizionario più ampio ed un tempo di indicizzazione maggiore rispetto agli altri per il fatto che vengono considerate tutte le parole presenti nei documenti e non viene applicata la fase di stemming.

	MAP	***P(10)***	***Rprec***
TF_IDF	0.2120	0.4800	0.2725
BM25	0.2126	0.4840	0.2705
BM25_stem	0.2108	0.4740	0.2740
TF_IDF_not	0.1875	0.4300	0.2460

Tabella 6*. Valori riassuntivi per ogni sistema.*

Vengono riportati in *Tabella 6* i valori riassuntivi delle misure considerate. Il valore della *MAP* è la media delle *AP* per ogni topic e questa mostra che il sistema *TF_IDF_not* ha un valore molto più basso rispetto agli altri, *BM25_stem* è poco sotto ai sistemi che utilizzano la stop list e la fase di stemming per l'indicizzazione. I valori per la misura *P(10)* sono una media su tutti i topic ed anche in questo caso i sistemi *TF_IDF* e *BM25* hanno valori più alti rispetto agli altri. Complessivamente quindi, questi due metodi hanno una *P(10)* migliore, ma nel caso occorra focalizzarsi in un ambito specifico, bisognerebbe eseguire delle ulteriori analisi per verificare le prestazioni su particolari topic e magari non considerare solamente i valori riassuntivi di tutti i topic.
I valori di *Rprec* penalizzano ancora il metodo *TF_IDF_not*, ma non premiano solamente i due modelli che fanno uso della stop list, questa misura quindi non definisce chiaramente quali siano gli IRS migliori.

4. Riferimenti

[1] Github: https://github.com/davidemartini/Information-Retrieval
[2] Terrier: http://terrier.org/
[3] Trec_eval: https://github.com/usnistgov/trec_eval
[4] os: https://docs.python.org/3/library/os.html?highlight=os#module-os
[5] matplotlib: https://matplotlib.org/
[6] statsmodel: https://www.statsmodels.org/stable/index.html
[7] numpy e scipy: https://docs.scipy.org/doc/
[8] PyCharm: https://www.jetbrains.com/pycharm/